CSC 478 – Executive Summary
June 5, 2017
Brandon Markwalder, William Martin, David Scroggins

Overview

Our analysis focused on the datasets generated by The National Longitudinal Study of Adolescent Health, which tracked American adolescents grades 7-12 from 1994 to 2008. After a thorough review of the dataset, we focused our analysis on classification of Wave 4 biomarkers (2008) for glucose levels, inflammation levels and lipid levels using a subset of the Wave 1 In-Home Questionnaire and Public Use Contextual Database (1994) as our predictors. In this summary, C_JOINT=1 indicates that an individual has diabetes (glucose biomarker), C_CRP=3 indicates an individual has "high" inflammation levels and C_JOINT2=1 indicates that an individual is hyperlipdemic (lipids biomarker).

The goal was to generate an overview of the association between self-reported social, behavioral and health factors a full 14 years prior to the collection of objective biomarker criteria. While the chronological distance between Wave 1 and Wave 4 along with the imbalanced distribution of most of the biomarkers presented a distinct challenge, we believed the longitudinal exploration and classification task presented the most exciting opportunity for analysis and knowledge discovery. The intent was to ask "What, if any, self-reported metrics are associated with long term health risks?" and "To what extent can these metrics be used to build a classification system for long term health risks?"

Model Selection

We chose to emphasize decision trees given their ability to classify mixed, complex data without requiring prior knowledge of the datasets features as well as their ability to generate easily interpretable results. By extension we chose to leverage the benefits of decision trees using the well-known ensemble technique, Random Forest. The hope was that in conjunction with the easy interpretability of decision tree results, Random Forests could provide additional interpretative depth and classification power. Finally, we considered K Nearest Neighbor as a classification model for making long term predictions for the biomarkers based on the Wave 1 predictors. KNN provides multiple benefits for this particular dataset, specifically 1) ease of interpretation, 2) relatively low computing time and 3) generally powerful predictive power. KNN is also well-suited for dealing with datasets containing a mixture of numeric and categorical variables. In all three instances we used the standard modules in sklearn: tree.DecisionTreeClassifier, ensemble.RandomForestClassifier and neighors.KNeighborsClassifier.[1]

[1] A handful of additional models were briefly considered. Naïve Bayesian Classifiers seemed potentially valid as an approach, particularly when applied to the large amount of categorical data contained in the dataset. In general, they did not perform well on the dataset, and given the model's lack of suitability for oversampling techniques (described below) were abandoned early. Adaboost was also considered as an approach for boosting our decision tree signals, however, the mediocre results pushed us in a different direction and were also abandoned. While

All model descriptions presume parameter optimization using GridSearchCV, which is detailed in the main report.

Model Performance

Initial results for both decision trees and random forests built on the glucose and lipids biomarkers suffered from low recall and precision when classifying the minority class. We were able to gain moderate improvements in recall using weighting, however, this could only be achieved by heavily weighting the minority class to the extent that we could not confidently interpret the results out of concern for overfitting. We opted instead for a well-known technique utilized in machine learning models for medical data: random oversampling. Utilizing a sklearn-compatible, third-party package, imblearn, we oversampled the minority classes using 10-fold cross-validation and 10-fold cross-validated predictions. In the case of both biomarkers we were able to significantly improve our recall and precision scores.

C_JOINT=1 DT		
precision	recall	f1-score
0.17	0.11	0.13

C_JOINT2=1 DT		
precision	recall	f1-score
0.03	0.01	0.02

C_JOINT=1 RF (Weighted)		
precision	recall	f1-score
0.11	0.42	0.17

C_JOINT2=1 RF		
precision	recall	f1-score
0	0	0

C_JOINT=1 DT (Oversampling)		
precision	recall	f1-score
0.58	0.47	0.52

C_JOINT2=1 DT (Oversampling)		
precision	recall	f1-score
0.51	0.58	0.54

C_JOINT=1 RF (Oversampling)		
precision	recall	f1-score
0.61	0.54	0.58

C_JOINT=2 RF (Oversampling)		
precision	recall	f1-score
0.56	0.47	0.51

For Inflammation and Immune Function we considered two variations of the biomarkers: the original three class version (“low,” “average,” “high”) and a two class version (“low to average”, “high,”) in both cases focusing on proper classification of “high” (C_CRP=3).

these models are not detailed in our report or summary, the results can be viewed in the script repositories linked in the main report’s Appendix.

3-Class DT		
precision	recall	f1-score
0.51	0.61	0.55

3-Class RF		
precision	recall	f1-score
0.45	0.83	0.58

2-Class DT		
precision	recall	f1-score
0.55	0.44	0.49

2-Class RF		
precision	recall	f1-score
0.64	0.25	0.36

2-Class DT (Oversampling)		
precision	recall	f1-score
0.57	0.57	0.57

2-Class RF (Oversampling)		
precision	recall	f1-score
0.62	0.76	0.68

Interpretation

In general, we focused our interpretation on the results of the oversampled decision trees using the random forest results as support for our interpretation. We emphasize that all interpretations should be approached cautiously. In most instances, we were able to general f1-scores higher than 0.5, which suggests we are identifying valid, non-random patterns. However, the results are not necessarily generalizable to the raw dataset, and should be viewed as tentative observations that can serve as the basis for deeper analysis. The granular details for this overview can be viewed in the main report under “Interpretation of Model Results.”

Two overarching themes unites the models, that of self-reporting on weight and median income levels. Individuals who report being overweight are consistently classified as diabetic and/or at risk for cardiovascular disease. While this is not revelatory, the presence of this feature in all three models is indicative that they are capturing meaningful data for each of the biomarkers. In addition, lower income was consistently aligned with a positive classification for a health risk. While this is a less obvious observation, socioeconomic status is heavily determinative of diet and by extension determinative of long term health outcomes.

For glucose levels, we must be careful interpreting the results as we do not have a clear mechanism to distinguish between Type 1 and Type 2, however, given the prevalence of Type 2 diabetes, we can cautiously assume that much of the captured data is reflective of individuals with this specific type. In addition to weight, median household income played an important classification roll pointing toward the importance of socioeconomic status for general health. Additionally, maternal involvement and education level play a secondary roll. Parental oversight and quantity of television watching also played minor roles. The latter two are more difficult to confidently interpret, but the general suggestion seems to be that students whose parents are better educated and more likely to monitor their activity have slightly better health outcomes when it comes to developing diabetes. These should not be received as absolute declarations, rather suggestions for deeper analysis to validate them as concrete patterns.

For inflammation and immune function scores, at least one interesting pattern emerged beyond the role of weight. For a small subset of student instances, access to dental care was predictive of high hsCRP scores and, by extension, cardiovascular risk. It is possible that oversampling amplified a small, previously hidden pattern in the data. We feel it is worthy of further exploration, though this data set is not likely to provide the proper context in which to do so. Among those not reporting weight issues, usage of illegal drugs was associated with a high hsCRP score. While illegal drug use might certainly lead to higher inflammation levels, it is important to remember that the self-reporting predates the biomarkers by 14 years. More likely is that this pattern captures a behavioral pattern that leads to long term health risks up to and including adverse risk for heart disease.

Beyond the role of self-reporting on weight, our lipid models produced little by way of useful results despite reasonable performance metrics. One possibility is that the role of genetics as a mechanism for hyperlipidemia simply makes it impossible to observe meaningful patterns beyond factors such as weight, which is a direct and known contributor. The results of the random forest model do indicate that median income and educational attainment play a roll, but it is difficult to directly interpret to what extent and in what manner they do so.

KNN

K-nearest Neighbors was applied to a normalized (range: 0-1) version of the raw dataset. As should be expected, KNN did not perform well on the raw data, especially given our utilization of the algorithm was to provide a generalizable classifier. KNN was unable to generate a recall score for either minority class C_JOINT=1 or C_JOINT2=1 higher than 0.1. Overall accuracy and recall for the majority classes was high due to the imbalanced distribution of the classes. KNN was optimized for k=3 and k=2 for C_JOINT and C_JOINT2 respectively. It is generally advisable to avoid even values of k, however, this represents to point at which recall performance for C_JOINT2=1 dropped precipitously.

KNN performed better on the 2-class version of the lipids biomarker (C_CRP), however the results were still suboptimal. K=16 provided the best results for the minority class, however, recall was still poor at 0.26 with precision of 0.5 and 0.34. While we can express some moderate confidence regarding our ability to extract promising patterns from the dataset using decision trees and random forests, we cannot confidently recommend KNN in this case as a long-range predictor for Wave 4 biomarkers based on Wave 1 self-reporting. However, we do present the possibility of optimizing KNN for stratified versions of the dataset in future analysis. For instance, application of KNN to a subset of instances with low median incomes (based on criteria from decision trees) as well as to a subset of instances reporting “slightly” or “very” overweight present themselves as potentially fruitful paths for analysis.

Wave 4 Glucose Preliminary Results

Given the difficulty of predicting diabetes based on Wave 1, we were interested in whether data in closer time proximity to the glucose biomarker would have higher predictive ability. So,

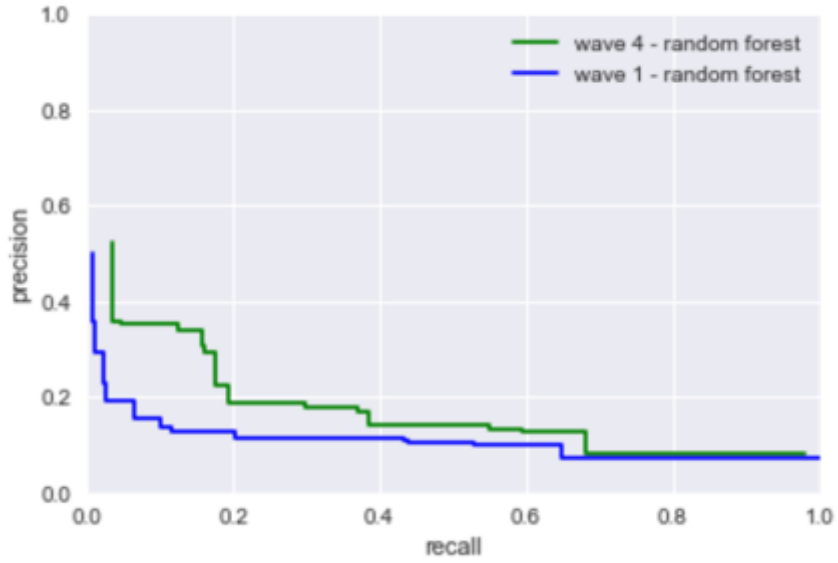
we ran a quick analysis into the Wave 4 survey data to see if we could improve our results. Since Wave 4 survey questions were coincident to the biomarkers measurement (and diabetes diagnosis), we were no longer building a forecasting model. Instead, we were building a quantitative inference (diabetes 'diagnosis') model based on qualitative data at a common point in time. Admittedly, the value of such a model is lower than that of a forecasting model, but could still pay off by confidently identifying low and high-risk individuals to optimize the usage of expensive biomarker measurement for a subset of the population.

Moving our analysis to another wave of data presented a scalability challenge. We manually selected important features from Wave 1 with a divide and conquer approach in which we examined the entire dataset, but we did not have time to do the same for Wave 4. There are 27 sections with a total of 919 questions to choose between. Furthermore, we manually distinguished categorical and numeric questions in Wave 1 (to tell which ones to dummy), another thing we did not have time for in Wave 4.

We applied heuristics to solve each of these problems. To distinguish questions that needed to be dummied, we counted unique responses for each question and treated questions with 10 or fewer unique responses as categorical (and dummied them). With this treatment applied, our 919 questions resulted in 2,649 independent variables. Considering the hardware we had available, we did not think training on this full set was a good approach. We managed the size of our predictors by measuring the correlation between each and our target feature, and then selected the top 500 to arrive at a data set roughly the same size as we used in Wave 1 analysis.

Initial analysis showed a very promising predictor (H4ID5D), with a .63 Pearson correlation to our target. However, the feature represented a survey question about whether the respondent had ever been diagnosed with diabetes. After excluding H4ID5D, we moved into modeling the data. We applied GridSearchCV to explore the precision recall tradeoff with Wave 4 data, but time restrictions prohibited applying oversampling techniques in Wave 4.

As expected, we did find higher predictive ability with Wave 4 data. In cross validation, our preferred model had an average recall of .55 and average precision of .14, slightly better than the comparable model in Wave 1 (recall = .43, precision = .11). We plotted the best random forest models along the precision-recall tradeoff from Wave 1 and Wave 4 to show the improvement we see with Wave 4 data.



The initial improvement is marginal, but gives further credibility to our assertion that predicting diabetes from the data at hand (especially without distinction between Type I and Type II) is a difficult task hampered by our inability to distinguish between diabetes type as well as account for potential genetic inheritance. We are eager to extend additional techniques to the Wave IV questionnaire as well as consider the other biomarkers in this context.